A Trusted Digital Repository based on the OAIS model with integrated management of access rights

Bernard Bel

Cultural Heritage on line – Trusted Digital Repositories & Trusted Professionals
Florence, 11-12 December 2012

A Trusted Digital Repository based on the OAIS model with integrated management of access rights

Bernard Bel
Laboratoire Parole et Langage (LPL)
CNRS - Aix-Marseille University
http://lpl-aix.fr
bernard.bel@lpl-aix.fr

Basic points

Laboratoire Parole et Langage (LPL), a speech research laboratory of the French Centre National de la Recherche Scientifique (CNRS), is in charge of an archive submission site called *Speech & Language Data Repository* (SLDR, www.sldr.org). The aim of SLDR is to preserve data eligible for (field or laboratory) speech/linguistic research or cultural heritage and facilitate its dissemination via non-commercial or shared licences. It is open to research and documentation projects worldwide. Currently (2012), SLDR is storing 262,000 documents dealing with 340 registered users from 46 countries.

SLDR is a generic service able to handle any type of item (information package). Resource pooling is constructed on an interoperable system involving two major computing centres (CINES and CC-IN2P3) in a joint project coordinated by the Adonis VLRI, currently the French branch of DARIAH (www.dariah.eu).

SLDR for speech, and CNRTL (www.cnrtl.fr) for text, are the resources centres from which a French sub-network of CLARIN (www.clarin.eu) centres is being built: the ORTOLANG project (www.ortolang.fr/english) associated with the CORPUS VLRI (www.corpus-ir.fr). This project is focusing on interoperability following CLARIN guidelines with respect to metadata (DC OLAC, CMDI…), persistent identifiers and controlled vocabulary (ISOcat).

Background of the project

In 2006 the office of the Social Science and Humanities department at the French *Centre national de la recherche scientifique* (CNRS) issued a call for projects aiming at the creation of digital data repositories for speech research. Two projects were selected under the label *Centre de Ressources pour la Description de l'Oral* (CRDO): CRDO-Aix and CRDO-Paris supported by *Laboratoire parole & langage* (LPL) and *Langues et civilisations à tradition orale* (LACITO) respectively. While CRDO-Paris

mostly replicated the design of the existing LACITO archive [1], CRDO-Aix was built from scratch after a comparative study of existing data repositories [2].

In 2008, TGE Adonis (www.tge-adonis.fr) was commissioned to promote the long-term preservation and sharing of oral resources in social sciences and the humanities. Its pilot project involved the two branches of CRDO as submission sites connected with major computing centres: *Centre informatique national de l'enseignement supérieur* (CINES, www.cines.fr) for long-term preservation and *Centre de calcul de l'Institut national de physique nucléaire et de physique des particules* (CC-IN2P3, cc.in2p3.fr) for data dissemination [3]. Following the example of spatial agencies, managers opted for the *Open Archival Information System* (OAIS) promoted by the *Consultative Committee for Space Data Systems*.

CRDO-Paris and CRDO-Aix became fully operational in Summer 2010 as their long-term preservation modules switched to the 'production' mode. At the term of this experimental phase in 2011, partners were instructed by INSHS (the social science and humanities institute of CNRS) to give up the 'CRDO' acronym. Thereafter, CRDO-Aix was renamed *Speech & Language Data Repository* (SLDR, www.sldr.org) without modifying its operational process.

A comprehensive implementation of the OAIS model

The *Open Archival Information System* (OAIS) is ISO 14721 standard. It was imposed by institutions heading the pilot project to comply with formal agreements between the French National Archive, CINES and CNRS. The actual implementation is addressing specific features of oral/linguistic resources, notably:

- The diversity of file formats: sound/video, all signals associated with speech/singing, pictures, texts, tables etc.
- Secondary data (annotations etc.) and metadata are mutable and extensible.
- Multilingual support for descriptive metadata, extra-European scripts, transliteration/annotation standards (IPA etc.)

Preserving data is the primary requirement of long-term preservation. Further, data should be eligible for reuse after an unspecified period of time (typically more than 30 years). To this effect, the project is relying on an institutional archive (CINES) rather than a consortium of computing centres. CINES is beneficiary of the *Data Seal of Approval* (sldr.org/wiki/DSA) as a first step towards certification. Its commitment is threefold: (1) preserving data and its associated metadata; (2) preserving access right information; (3) preserving the usability of data which is achieved by migrating file formats (without loss of data) when these are becoming obsolete.

Procedures for submitting and retrieving information packages to/from SLDR are illustrated figure 1. Data is first sent to CINES in Montpellier (left part) and assessed for correctness with respect to long-term preservation: compliance with structural specifications for the *Submission Information Package* (SIP), acceptance of file formats and consistency of file contents. This automated check-up is purely technical as the informational content (scientific correctness and relevance) remains under the responsability of the submission site.

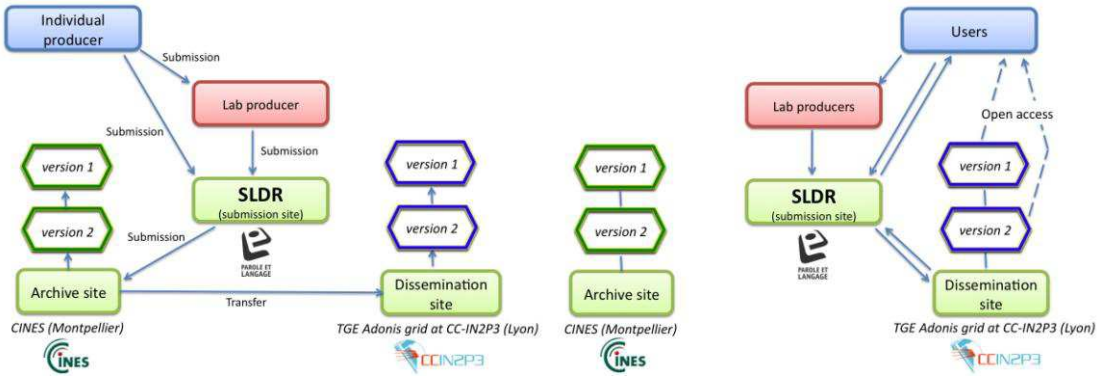

Fig . 1: Submitting and retrieving data to/from SLDR.

If the SIP is accepted it is completed with archival metadata and assigned an an ARK (*Archival Resource Key*) identifier to be stored as an *Archival Information Package* (AIP). This AIP is forwarded along with dissemination information to the dissemination site (CC-IN2P3) in Lyon which is running a Fedora Commons (www.fedora-commons.org) repository.

Access to data is illustrated on the right part of figure 1. Users may get direct access to datastreams on the dissemination site when these are in open access. Otherwise they must proceed to a (non-commercial) transaction with SLDR: identify themselves as a registered user belonging to a group authorized for the particular resource or file, check SLDR licence (and optional ones) for approval, following which SLDR will set up a 'channel' for downloading the datastream(s) from CC-IN2P3. In neither case CINES is involved in the dissemination of data.

SLDR is able to reconstruct the source data and metadata of an item from datastreams stored at the dissemination site. This makes it possible to delete items on the submission site after their successful submission for long-term (or medium-term) preservation.

Dealing with mutable/extendible data

In a given item there are bits of information that remain persistent (e.g. primary data of sound/video resources) whereas others (such as descriptive metadata and access rights settings) may need to be updated more frequently. An institutional archive preserves all versions of successfully submitted information packages. It would be detrimental to upoload a new version of the entire package every time a small amount of its descriptive data or access rights have been modified.

This problem has been solved in the OAIS framework thanks to a sophisticated structure of Submission Information Packages. Each SIP contains two folders: one which is dedicated to persistent data and the other one to "documentary files". It is possible to submit a new version either for the entire package or only its documentary files and/or access rights settings. The latter is called "metadata updating".

Item segmentation, ARK identifiers and PIDs

Technical constraints on the archive and dissemination sites imply limitations of the size of Archival Information Packages. Currently, packages may not exceed 40 Gbytes / 30,000 files. Therefore, large items need to be segmented to several AIPs, a segmentation that remains invisible to users. For example, *The Open ANC* (sldr000770) containing more than 60,000 files is segmented to seven AIPs. Each AIP is assigned an ARK (*Archival Resource Key*) identifier. There is no global resolution mechanism for ARK identifiers. We implemented a local name mapping authority, e.g: http://sldr.org/ark:/87895/1.4-183706 points to segment 2 of *The Open ANC*.

A finer grain of identification is required that does not rely on AIPs. To this effect, a persistent identifier (PID) is assigned to each item — yet not every AIP. We use the Handle system (www.handle.net) and construct PIDs with comprehensive syntax, e.g.: hdl:11041/BeQuali-000532 (case insensitive) points to item *BeQuali-000532* whose canonic URL is http://sldr.org/BeQuali-000532 (case insensitive). Some PIDs need to point directly to a document whereas others may point to a descriptor of (a set of) document(s) which may be machine-readable (e.g. metadata in a XML format) or human-readable (e.g. a web page). Currently, in SLDR, a PID pointing to an item displays its descriptive page but suffixes will be used to retrieve metadata in OAI_DC, OLAC, CMDI etc. formats.

PIDs are also assigned to each file contained in each version of an item, e.g.: hdl:11041/swedia-000788_v1_f388 points to file index 388 of version 1 of item *swedia-000788*. If the version number is not specified then the latest version is fetched. We are planning to implement optional alternate PIDs in which file indexes may be replaced with identifiers derived from original file names so that the same document remains traceable should its index be modified accross versions.

Systematic use of PIDs is an important topic under debate in *Virtual Competency Centres* VCC1-VCC3 of the DARIAH network (www.dariah.eu) and the APARSEN Network of Excellence (www.aparsen.eu). Therefore, we avoid implementing early options that might turn out obsolete once a broad consensus has been reached on guidelines conducive to better interoperability.

Integrated management of access rights

France takes advantage from a significant advance on archive law, owing to its *Code du patrimoine* (the Heritage Code) clarifying the status of "public archive" with a set of formal rules regulating the access to archived documents (Act of 15 July 2008, articles L213 1-5). This framework implies a radical change of practice with respect to the long-term preservation and sharing of digital data, as any public archive shall be immediately in open access with the exception of 24 derogations applicable to certain categories of documents.

According to Art. L213-5, *"Any administration [...] is compelled to give reasons for denying access to a public archive."* This implies that when attempting to download a document under restricted access, users should be given the reason for denial and the date of its future public release. This process is illustrated on figure 3. A sound corpus

is publicly available in 'low' resolution (AAC format) whereas the same files in 'high' resolution (WAV format) are shared with identified scholars exclusively. This procedure is compliant with research participants' agreement that anyone may listen to their speech production while remaining cautious about possible alterations and misuse of source files by unidentified persons.

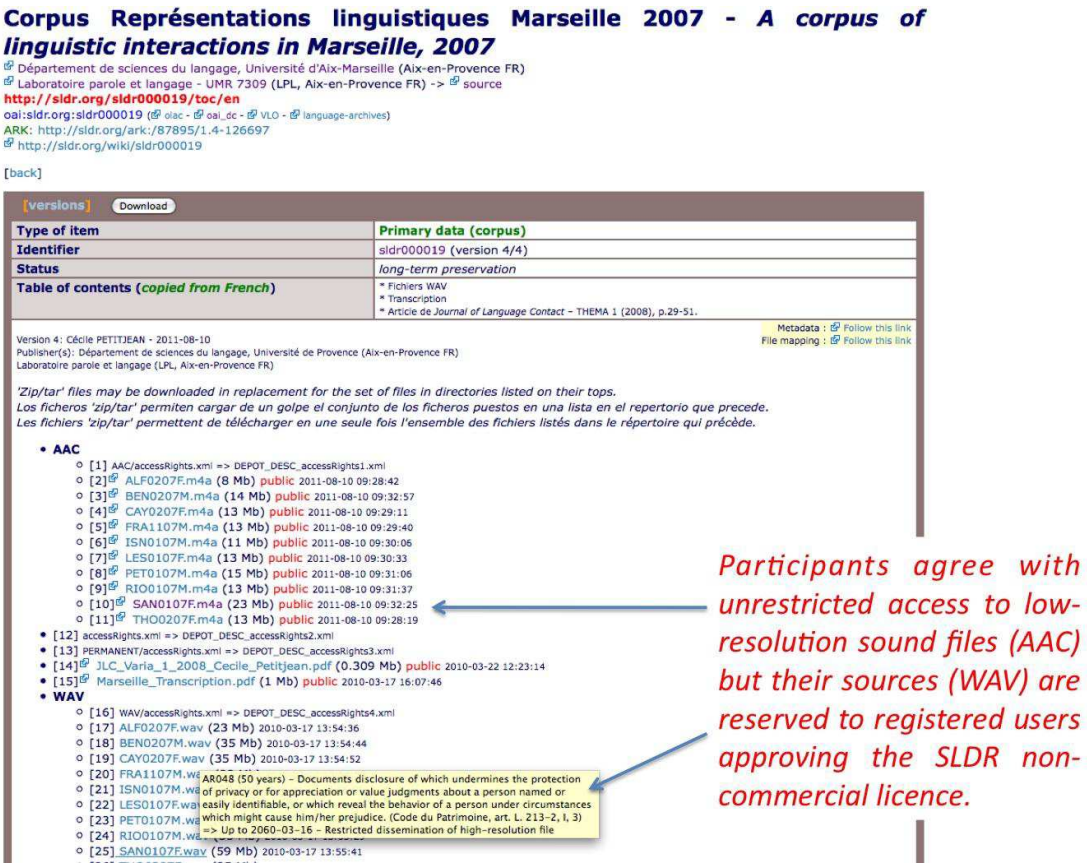

Fig. 2: File sharing in 'high' and 'low' resolutions

In this example, moving the mouse over a restricted link displays the AR048 derogation to public access: "protection of private data", a frequent property of ethnographic and oral/linguistic data. This warning is displayed in the navigation language chosen by the user and it contains the date at which the document is expected to become open-accessible. In its daily set-up, SLDR brodcasts messages to administrators notifying them that the status of an item or document needs to be changed from restricted to public (or the other way around).

Clicking the link prompts the user to sign in on the site, check her/his status, verify whether access is granted to this document and request the user to approve licence(s) attached to this item. This whole process is stored in memory so that it will be bypassed when accessing similar documents during the same session. In this way, complex queries may be performed on multiple documents, multiple items and later multiple sites once Single Sign-On has been implemented among partner repositories.

Access right conditions must be made visible in OAI-PMH metadata as shown figure 4. Usage of controlled vocabulary borrowed from the *info:eu-repo* namespace is compliant with DRIVER (www.driver-repository.eu) and OpenAIRE (www.openaire.eu) portals.

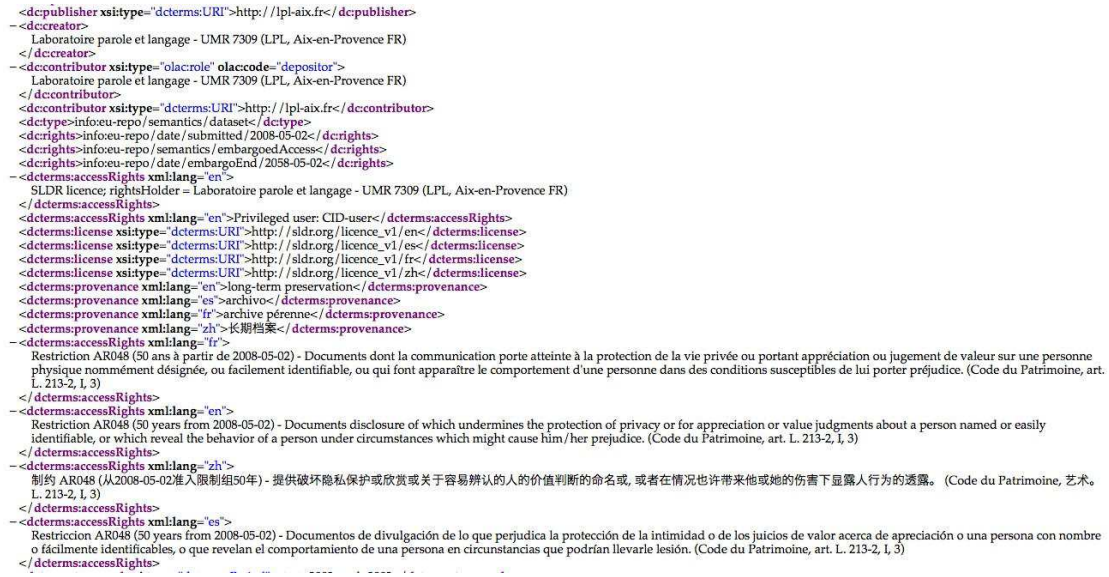

Fig. 3: Access right information in the OLAC DC for item hdl:11041/sldr000027

Shared licences

Sharing an item of SLDR within an institution is made possible via a licence that may be commercial (a group purchase) or non-commercial. Example of non-commercial licence: *Buckeye Corpus of Conversational Speech* distributed by the Ohio State University, hdl:11041/sldr000776. The following steps are required:

- The institution is referenced on SLDR site;
- The user is registred on SLDR site and affiliated with this institution;
- A copy of the resource is disseminated by SLDR;
- SLDR owns documents proving that the licence is valid for the institution.

Conclusion

We hope that policy makers will go on supporting the construction of very large research infrastructures for social sciences and humanities in association with networks such as CLARIN and DARIAH in Europe. Their commitment is crucial in response to a strong demand for cooperative resource development and sharing.

References

[1] Michailovsky, B.; Michaud, A.; Guillaume, S. (2011). A simple architecture for the fine-grained documentation of endangered languages: the LACITO multimedia archive. *International Conference on Speech Database and Assessments* (Oriental COCOSDA 2011), Hsinchu: Taiwan. halshs.archives-ouvertes.fr/halshs-00620893

[2] Bel, B.; Blache, P. (2006). Le Centre de Ressources pour la Description de l'Oral (CRDO). *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-*Provence (TIPA), 25, pp. 13-18. hal.archives-ouvertes.fr/hal-00142931

[3] Bärring, O. (2008). Hosting of IT services and data for Human and Social Sciences in France. A preliminary study for TGE Adonis (Contract Nr K1432). www.sldr.org/docs/admin/RapportBarring.pdf